DOCUMENT RESUME

ED 455 289 TM 033 143

AUTHOR Rivera, Charlene; Stansfield, Charles W.
TITLE The Effects of Linguistic Simplification of Science Test Items on Performance of Limited English Proficient and Monolingual English-Speaking Students.
SPONS AGENCY Delaware State Dept. of Education, Dover.
PUB DATE 2001-04-12
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Competence; *Difficulty Level; *Elementary School Students; English (Second Language); Intermediate Grades; *Limited English Speaking; Science Tests; State Programs; *Test Items; Testing Programs
IDENTIFIERS *Simplification (Language); Testlets

ABSTRACT

This study examined the effects of linguistic simplification of fourth and sixth grade science test items on a state assessment on the achievement of English language learners, those of limited English proficiency (LEP). At each grade level, 4 parallel 10-item testlets were included on an operational statewide assessment. On each testlet, items differed only in that on one testlet they were linguistically simplified, while on the other, the standard wording was used. The testlets were randomly assigned to LEP and non-LEP students through the spiraling of test booklets. For non-LEP students, in four T-test analyses of the differences in means for each corresponding testlet, three of the mean score comparisons were not significantly different, and the fourth showed the regular version to be slightly easier than the simplified. Analysis of variance, followed by pairwise comparisons across the two conditions, showed no significant differences in the scores of non-LEP students for any of the comparisons across the two-item types. Among the 40 items administered in both regular and simplified format, item difficulty did not vary consistently in favor of either format. Qualitative analysis of items that displayed significant differences in P values was not informative, since the differences were typically very small. For LEP students, there was one significant difference in student means, and it favored the regular version. However, the LEP student analyses lacked statistical power due to the small sample size and the low reliability of testlets for this sample. The results of the study show that linguistic simplification is not helpful to non-LEP students who receive it. Therefore, the results provide evidence that linguistic simplification is not a threat to score comparability. (Contains 25 references and 8 tables.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 455 289

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
C. Rivera
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

The Effects of Linguistic Simplification of Science Test Items on Performance of Limited English Proficient and Monolingual English-Speaking Students

Charlene Rivera
The George Washington University
Center for Equity and Excellence in Education

and

Charles W. Stansfield
Second Language Testing, Inc.

TM033143

Paper presented at the
Annual Meeting of the
American Educational Research Association
Seattle, Washington
April 12, 2001

Acknowledgement. We wish to express our appreciation to the Delaware Department of Education (DDOE) for having funded this study. At the DDOE, Drs. Wendy Roberts, Nancy Maihoff, and Liru Zhang were helpful and supportive, as was Harcourt Educational Measurement, Delaware's contractor for the Science test.

BEST COPY AVAILABLE

The Effects of Linguistic Simplification of Science Test Items on Performance of Limited-English-Proficient and Monolingual English-Speaking Students

Abstract

The use of accommodations has been widely proposed as a means of including ELLs in state and district-wide assessments. However, very little experimental research has been done on specific accommodations to determine if they are a threat to score comparability and to determine their usefulness for LEP students. This study examined the effects of linguistic simplification of fourth and sixth grade science test items on a state assessment. At each grade level, four parallel 10-item testlets were included on an operational state-wide assessment. On each testlet items differed only in that on one testlet they were linguistically simplified, while on the other, the standard wording was used. The testlets were randomly assigned to LEP and non-LEP students through the spiraling of test booklets. For non-LEP students, in four t-Test analyses of the differences in means for each corresponding testlet, three of the mean score comparisons were not significantly different, and the fourth showed the regular version to be slightly easier than the simplified. ANOVA, followed by pairwise comparisons across the two conditions, showed no significant differences in the scores of non-LEP students for any of the comparisons across the two item types. Among the 40 items administered in both regular and simplified format, item difficulty did not vary consistently in favor of either format. Qualitative analysis of items that displayed significant differences in P values was not informative, since the differences were typically very small. For LEP students, there was one significant difference in student means, and it favored the regular version. However, the LEP student analyses lacked statistical power due to the small sample size and the low reliability of the testlets for this sample. The results of this study show that linguistic simplification is not helpful to non-LEP students who receive it. Therefore, the results provide evidence that linguistic simplification is not a threat to score comparability.

Introduction

In recent years, there has been much discussion about how to best assess the school achievement of English language learners (ELLs) or limited English proficient (LEP) students[1]. Two problems faced by those charged with setting inclusion and accommodation policies for state assessment programs designed for system level monitoring and accountability are: 1) the lack of research on the effects of accommodations generally (Shepard, Taylor, & Betebenner, 1998); and 2) the lack of research on how specific accommodations address the linguistic needs of ELLs. This paper reports on a study of one accommodation, simplified English, used in the context of the Delaware state assessment program.

Currently, in the Delaware state assessment system, the accommodation of simplifying or paraphrasing test directions or questions is considered a Condition 3 accommodation. Condition 3 accommodations are not included in the school and district means because there is concern that students who receive such accommodations will be significantly advantaged. However, since the practice of linguistically simplifying test items is a promising accommodation strategy for ELLs, an experimental study was designed to assess its effects for English language learners and monolingual English speakers.[2]

The results of the study described in this article should contribute to an understanding of the effects of linguistically simplifying test items on test scores of both ELLs and monolingual English speakers, at least in the context of elementary science assessments in Delaware. The study findings may also have implications for the use of the linguistic simplification accommodation in subject areas other than science.

Review of Literature on LEP Student Accommodations

Historical overview. State assessments, like standards-based education, are closely linked to accountability. It is widely believed that school achievement will improve if education systems identify what is to be learned, and then assess that material to determine the effectiveness of instruction (CED, 2000). However, concern has been raised about the degree to which standards and accountability systems will include language minority students generally and limited English proficient students or English language learners specifically (Rivera, & LaCelle-Peterson, 1993; LaCelle-Peterson & Rivera, 1994). In

---

[1] In this article, Limited English Proficient (LEP) students and English Language Learners are used interchangeably. LEP is the term used in the Elementary Secondary Education Act to refer to students whose first language is not English and who are designated eligible to receive English as a second language and bilingual services. The term English language learner (ELL) is used to refer to the same students but focuses "on what students are accomplishing, rather than on any temporary limitation they face" (LaCelle-Peterson and Rivera, 1993, p.55).

[2] In June 1999, Delaware issued a request for proposals for research and development on accommodations for LEP students. The authors responded to that request with a proposal to carry out a study of the accommodation of linguistically simplifying science test items and in carrying out an experimental study of the effect of simplifying science test items.

1994, Rivera and Vincent (1997) conducted a survey of state policies concerning the inclusion or exemption of ELLs during the 1993-94 school year. In response to a questionnaire sent to state education agencies, they found that 44 of 48 states with state assessment programs permitted ELLs to be excused from one or more state assessments. In 27 of the 44 states, ELLs as a group were routinely exempted from participation in the state assessment program. Rivera and Vincent concluded that states needed to focus on including greater numbers of ELLs in the state assessment program, if they were to be expected to attain the same high performance standards anticipated for monolingual English general education students. To encourage states to hold ELLs to the same standards, Rivera and Vincent (1997) recommended the judicious use of accommodations in testing and the development of alternative test options. They also recommended that states collect data and carry out studies to evaluate the impact of various types of interventions on LEP student achievement.

While Rivera and Vincent (1997) were conducting their study, an independent journalist with support from the MacArthur Foundation conducted an investigation of the testing practices of the 14 largest school districts in the United States (Zlatos, 1994). He found that exemption of the least able students (students with disabilities, ELLs, and low achievers) was a common practice, and that there was substantial variation in the percentage of students included in the testing program. For example, he found that 87% of students were tested in Philadelphia, 76% in New York City, 70% in Washington DC, and 66% in Boston. The conclusion drawn by Zlatos was that test scores are used as comparative evidence of the quality of schools, without disclosing that the least able students are regularly exempted from participation in the assessments. Zlatos findings clearly suggest that learning disabled and limited-English-proficient students cannot benefit from the standards-based movement unless they are included and reported on in state and district assessments and accountability systems.

A subsequent study of state inclusion and accommodation policies for ELLs in the 1998-1999 school year (Rivera, Stansfield, Scialdone, and Sharkey, 2000) showed states were generally trying out various accommodations with ELLs. However, the evidence suggests that most states utilize accommodations designed for students with disabilities rather than accommodations designed with the linguistic needs of ELLs in mind. In the book chapter, "Leveling the Playing Field for English Language Learners", Rivera and Stansfield (1998) propose criteria and outline procedures for making decisions about the inclusion and accommodation of ELLs in formal assessment programs.

Legal and legislative overview. About the time Zlatos was conducting his investigation, the Elementary Secondary Education Act of 1994, known as the Improving America's Schools Act (IASA) was reauthorized. It contains requirements that *all* "students" reach challenging content and performance standards, and be included in state assessment systems in at least mathematics and reading or language arts.[3] Specifically, the law requires that LEP students be assessed annually "to the extent practicable in the language

[3] The rationale for linking standards and assessments is that inclusion of all students in the assessment system will influence what is taught and how it is taught and provide educators with feedback to guide instructional practices.

and form most likely to yield accurate and reliable information on what such students know and can do, to determine such students' mastery of skills in subjects other than English." (IASA, Section 1111(b)(3)(F)(iii). Guidance from the US Department of Education for reviewing evidence of final assessments, discusses strategies for how to best assess LEP students. Specifically, it discusses the use of accommodations and mentions the use of "an assessment that has been stripped of non-essential language complexity" to assess an LEP student (USDE, 1999, p. 15).

Clearly, there is a need for evidence about the appropriateness of specific accommodations for ELLs. The major concern is their effect on score comparability, reliability, and validity. This concern is also voiced in the IASA legislation, which states that all assessments systems used for Title I programs "must be valid and reliable and be consistent with relevant, nationally recognized professional standards" (OESE, October 1, 1996). According to the Draft 1996 IASA Guidance on Assessments, "assessment measures that do not meet these requirements may be included as one of the multiple measures [of adequate yearly progress] if the State includes in its State plan sufficient information regarding the State's efforts to validate the measures and to report the results of those validation studies." (OESE, p.15, 1996). A requirement of IASA is that states create final assessment systems that are inclusive of all students by the 2001 school year.

However, since there are a very limited number of in-depth studies evaluating the effects of accommodations on ELL performance, it continues to be critical to study the effectiveness of specific accommodations for ELLs. In addition to the federal legislative impetus of IASA, there have been many calls from the education and measurement communities for research to identify appropriate, valid and reliable accommodations for ELLs (e.g., *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999); the *Position Statement of the American Educational Research Association Concerning High Stakes Testing in Pre-K-12 Education* (AERA, 2000); the Teachers of English to Speakers of Other Languages *Position Paper on Assessment and Accountability for ESEA Reauthorization* (TESOL, 2000); and the USED Office for Civil Rights' guidance on the use of tests to make high stakes decisions on students (USED OCR, 2000). While research on accommodations for ELLs has begun to be reported on at conferences and to appear in the literature (Stancavage, Allen, & Godlewski, 1996; Olson, & Goldstein, 1997; Abedi, et. al. 2000), studies involving accommodations rarely involve an experimental research design, making it difficult to determine the effects of accommodations on reliability, validity, and score comparability (Shepard, Taylor, & Betebenner, 1998).

Relevant research. We are aware of only one study of linguistic simplification as an accommodation for ELLs. Abedi conducted this study with mathematics items used on the National Assessment of Educational Progress (NAEP). In the study, test booklets containing either a Spanish version, a simplified English version, or original NAEP math items (in unsimplified English) were randomly administered to 1400 8th grade language or national-origin minority students in southern California middle schools (Abedi, 1997). Only Hispanic students received the Spanish version. The simplified items were rewritten by content experts in linguistics and math at the Center for Research on

Evaluation Standards and Student Testing (CRESST). The analyses indicated that both LEP and non-LEP (fully-English-proficient [FEP]) students performed best on the simplified version, and worst on the Spanish version. While LEP and non-LEP students performed significantly better on the simplified items, significant differences in item difficulty were obtained on only 34% of the simplified items. Abedi concluded that linguistic clarification of math items might be beneficial to all students. He also noted that other factors, such as length of time in US, English proficiency, reading competency, and prior math instruction, also had a significant effect on scores.

In a less formal study, the Colorado Department of Education in 1998 experimented with different versions of a state assessment for LEP students. They administered a simplified version, a version with an English glossary containing definitions of non-technical words, and a Spanish version to LEP students. They found favorable responses to the simplified version and the version containing an English glossary (Kiplinger, V. personal communication, 1999).

Implications. The results of the Abedi study and the effort in Colorado provide evidence that linguistic simplification of items may have utility as an accommodation for ELLs and language minority students taking formal assessments. However, more research is needed to attain a full understanding of the effect of linguistic simplification as a test accommodation on ELL scores and the scores on regular (monolingual general education) students. Only through a full understanding of the effects of simplification will it be possible to determine if it should be viewed as a threat to validity and/or score comparability.

Statement of the problem/research hypotheses

The null forms of the research hypotheses explored in this study are:

1. The mean raw score for grade 4 and 6 LEP students on linguistically simplified science items will not be significantly greater than that of LEP students taking the standard version of the same items on the DSTP Science test.

2. The mean raw score for grade 4 and 6 monolingual English speaking students on linguistically simplified science items will not be significantly greater than that of similar students taking the regular version of the same items on the DSTP Science test.

3. The difficulty of linguistically simplified science items will not be significantly different from difficulty of regular items for LEP students in grades 4 and 6.

4. The difficulty of linguistically simplified science items will not be significantly different from difficulty of regular items for monolingual English speaking students in grades 4 and 6.

Instrumentation

The Delaware Student Testing Program (DSTP) is based on approved content standards for the teaching of English language arts, mathematics, science, and social studies. State assessments in English language arts and mathematics were administered for the first time in the spring of 1998 and again in the spring of 1999 to students in grades 3, 5, 8, and 10. Assessments in science and social studies for grades 4 and 6 were field tested in the fall of 1999. The results of the field-testing were used to assemble the final forms of the tests and the first operational administration occurred in the fall of 2000, when this study was conducted.

In determining which tests to simplify for this study, the researchers examined the sample items that are available on the DDOE web site. An examination of the sample items in math, science, and social studies, indicated that the science items might benefit most from linguistic simplification. This judgment was confirmed during a more detailed examination of secure items on the math, science, and social studies tests during a visit to the DDOE. While all these tests contain certain items that can be simplified in terms of the level of language employed, the language load in the math test is somewhat less than in the science test, and the language of the social studies test is more intimately intertwined with the expression of concepts presented and measured on the instrument. Thus, the Science test was chosen for this study.

Forms of the science assessment at both grade levels consist of 50 items. Thirty-two are four-option multiple-choice items and 18 are short response items. The multiple-choice items are scored dichotomously, as either right or wrong (0 or 1 point for each item). The 18 short response items are scored on a 0-2 scale, with 0 generally representing an inappropriate response or no response, 1 indicating a partially correct response, and 2 representing a fully correct response. To earn a 2, the student must generally demonstrate knowledge of the correct answer and explain why the answer is correct. The latter aspect demonstrates conceptual understanding.

Research Design and Methodology

The original plan was to carry out this study using full-length operational tests in the DSTP. However, since the effects of linguistic simplification were unknown, it was feared that those students who took the simplified version would have an unfair advantage. Therefore, it was subsequently decided that the study should be conducted with the field test items embedded in the operational tests.

Each DSTP Science assessment consists of four forms at each of the four grade levels included within the program. Each form contains a combination of 40 operational items and 10 field test items. For purposes of this study, two additional forms were created for two grade levels. These additional forms were identical to two of the regular forms, except that the 10 field test items were simplified. The six operational forms administered at each grade level in the fall of 2000 are listed in Table 1.

TABLE 1
DSTP Science Assessment Forms and Treatment

Form	Treatment
A	Regular
B	Regular
C	Regular
D	Regular
E	Simplified
F	Simplified

All 4th and 6th grade students in Delaware participated in the study, since data was collected on all students, regardless of which form they took. However, our analyses were based only on the test performance of students who took forms C through F. Since there are almost 9,000 students at each of these grade levels in Delaware, each form was administered to a sample of almost 1500 students. The forms were randomly assigned to students through a spiraling procedure, so that in each classroom all six forms were used. Thus, each form was taken by approximately 1/6 of all tested students in the state at that grade level.

Forms A and B did not contain any items that were involved in this study. Forms C and D each contained 10 field test items as written, reviewed and revised by Delaware teachers, and then reviewed and edited by test development staff at Harcourt Educational Measurement, the testing contractor used by Delaware. Each item underwent multiple iterations or review and revision both in Delaware and at Harcourt. The 10 field test items on each form consisted of 8 multiple-choice (MC) and 2 short answer (SA) items. Thus, there were 20 items involved in the study at each grade level. These 20 items were divided into two testlets of equal length, with each testlet being randomly assigned to monolingual English speaking and Limited English Proficient students. In Delaware, MC items are scored as right or wrong, while the SA items are scored on a 3-point scale with 0 to 2 points being awarded for each item. All items are based on the state content standards for Science. The grade 4 items assess mastery of the grades K-3 standards, while the grade 6 items assess mastery of the grades 4-5 standards.

The forms were administered to all eligible students. The sample included regular monolingual English-speaking students, an unknown number of fully-English-proficient bilingual students, and limited-English-proficient students who had been in Delaware schools for more than one year. Students who have been in the system for less than one year are eligible for exemption from participation in the DSTP by state policy. Although a variety of accommodations are allowed, many LEP students who are tested in Delaware take the tests without accommodations.

The state also collects data and maintains a database on all students. This database includes information on a variety of background variables, including school, district, sex, race or ethnicity, learning or other disability status, Title I status, income

status, LEP status, migrant status, and category of accommodation received (permitting aggregation or disaggregation of test scores).

Test simplification. Once DDOE and Harcourt agreed on the final version of the field-test items they were sent by Harcourt via overnight courier to The George Washington University Center for Equity and Excellence in Education in May 2000. The next day, project staff and consultants including two middle school science teachers, three applied linguists, and an English as a second language (ESL) test developer met to review the field-test items.[4] The intent of the simplification process was to further clarify for LEP examinees the task or the context for each item and to reduce its reading difficulty level. The simplification of the items was completed in one day, and the revisions were sent via overnight mail to Harcourt and the DDOE the following day. The tests were administered between October 10 and October 19, 2000.

The design of the study made it possible to examine a number of issues related to the Science items. These issues relate to the effects of linguistic simplification on LEP and regular students' test scores. The effects could be determined at the level of the testlet (mean score per 10-item testlet by language proficiency status [LEP or non-LEP]), and at the level of individual items as well (P values by language proficiency status). The study was replicated at two grade levels. Thus, trends and effects at one grade level could be examined for consistency at the other. The design also made it possible to compare the psychometric characteristics (i.e., reliability) of the testlets by type of item (simplified vs. regular) for each group of examinees at each grade level.

In order to determine whether there were significant differences in group means across test versions (regular or simplified), t-tests for independent samples were conducted at each grade level. Analysis of variance was used to further explore the data in a way that involved all forms at each grade level simultaneously. The Duncan (1955) and Scheffé (1953) procedures were used to make pairwise comparisons when a significant overall F resulted from the analysis of variance. Due to the presence of the four short answer (SA) items scored 0, 1, 2, analysis of variance was also used to compute reliability coefficients for each of the 10-item testlets by form within grade.

---

[4] The science teachers were Jim Egenreider and Ray Leonard of Fairfax County (VA) Public Schools. The applied linguists included the authors and Dr. Judith Gonzalez of The George Washington University Center for Equity and Excellence in Education. The ESL test development specialist was Dr. John Miles of the TOEFL test development area at Educational Testing Service. Miles also coauthored with us a training manual on linguistic simplification of test items that was developed as part of this study (2000).

Results

A total of 11,306 non-LEP students took one of the eight[5] forms compared in this study. The number of students taking each form was approximately 1400. A total of 109 LEP students took one form of the test. Because this number was divided across eight forms, the number taking each test form was disappointingly small, and ranged from 6 to 23 students per form. Because the LEP samples were too small to provide generalizeable results, it is not appropriate to try to interpret the findings for the LEP group that participated in this study. However, the sample sizes for the non-LEP group were more than adequate for analysis and interpretation.

Item scores, either 0 or 1 for the multiple choice items and 0, 1, 2 for the short answer items, were summed across the 10 regular or simplified items in order to develop a total score for each examinee. Comparison of means on each type of item (regular or simplified) were made within a grade level for both the non-LEP and LEP groups using both t-tests and analysis of variance.

t-Test Comparisons

Using t-tests, within each grade level, the mean of form C was compared to the mean of form E, and the mean of form D was compared to the mean of form F for both non-LEP and LEP examinees. The results are displayed in Tables 2 and 3 below.[6]

Non-LEP Examinees. Table 2 shows the mean scores on the regular and simplified items for non-LEP examinees. The table shows that the difference in 4th grade students' mean scores on forms D and F was not significant. The difference in mean scores on forms C and E was significant ($p<.05$). The mean score for the sum of the regular items was 6.83, which was significantly greater than the mean of the sum of the simplified items (6.57). However, this difference favoring the regular version, is very small, amounting to only 2.5% of the range of possible scores on the testlet.

Table 2 also shows mean score comparisons for the 6th grade students on forms C and E and forms D and F. For 6th grade students, the difference between the mean scores in these two comparisons was not significant at the $p<.05$ level, despite the large sample. This finding suggests that there is no advantage for regular English-speaking students who took simplified items when compared to regular English-speaking students who took the regular items. Overall, in three of four comparisons among the Non-LEP students, no significant difference in performance was found. In the fourth comparison (Forms C with E at grade 4), only a very slight difference was found and it favored the regular version.

---

[5] Although six test forms were administered at each grade level, only four contained items that were compared in this study. Therefore, a total of eight forms were compared over the two grade levels.

[6] Dr. John Martois, an independent statistical consultant, carried out all data preparation and analysis for this study.

TABLE 2
t-Test for the Difference Between Mean Scores on the Regular and 10 Simplified Field-test Items for Grades 4 and 6 Non-LEP Examinees

Grade 4	Form	Type	*n*	Mean Score	*SD*	*t*	*df*	*p*
	C	Regular	1430	6.83	2.64			
						2.63	2840	.009
	E	Simplified	1412	6.57	2.69			
	D	Regular	1426	6.57	2.65			
						0.42	2840	.676
	F	Simplified	1416	6.61	2.65			
Grade 6								
	C	Regular	1415	4.86	2.37			
						0.99	2782	.322
	E	Simplified	1368	4.95	2.24			
	D	Regular	1416	6.44	2.60			
						1.66	2837	.096
	F	Simplified	1423	6.61	2.64			

LEP Examinees. Table 3 shows the mean scores on the regular and simplified items for LEP students in 4th and 6th grades. The very small sample size of the LEP groups (Ns between 6 and 23) strongly suggests that the findings for LEP students cannot be generalized. Among the four comparisons made, only one was statistically significant, and it favored the group receiving the regular items. There were no significant differences at the 4th grade level. For 6th grade students, the difference between the mean score on forms C and E was significant ($p<.05$). The mean for the regular items was 4.00, which was significantly greater than the mean (2.11) for the simplified version of these items. In the other 6th grade comparison, the difference in means on forms D and F was not significant.

TABLE 3
t-Test for the Difference Between Mean Scores on the Regular and 10 Simplified Field-test Items for Grades 4 and 6 LEP Examinees

Grade	Form	Type	*n*	Mean Score	*SD*	*t*	*df*	*p*
4	C	Regular	15	4.67	1.91	0.42	31	.677
	E	Simplified	18	4.33	2.52			
	D	Regular	23	3.48	1.89	1.52	37	.137
	F	Simplified	16	4.38	1.71			
6								
	C	Regular	9	4.00	1.50			
						2.88	16	.011
	E	Simplified	9	2.11	1.27			
	D	Regular	13	3.23	2.45			
						1.09	17	.289
	F	Simplified	6	2.00	1.79			

Analysis of Variance

Non-LEP Examinees. In order to further analyze the data for any differences between means, a one-way analysis of variance was computed at each of the two grade levels for non-LEP students (see mean scores shown in table 2). The independent variable, test forms, included 4 forms of the test, C, D, E, and F. The dependent variable was an examinee's score on the 10 items. Forms C and D consisted of regular items while forms E and F contained simplified items. The overall F ratio at both 4th and 6th grade levels was significant at the $p<.05$ level, suggesting a slight difference in scores across the large sample of non-LEP students.

In order to determine which means differ significantly from the others, post hoc pairwise comparisons were made using Duncan's (1955) Multiple Range Test. At the 4th grade level, Duncan's procedure indicated that the mean for form C (with regular items) was significantly greater than the mean of the other three forms. Scheffé's (1953) more conservative procedure, which keeps the overall error rate at $p<.05$ for all comparisons, failed to find any significant differences between pairs of means.

Post hoc pairwise comparisons at the 6th grade level provided consistent results with both the Duncan and Scheffé procedures. The means of forms D and F were significantly greater than the means of forms C and E. However, forms D and F are alternate versions of the same 10-item testlet, and on these two versions no differential advantage was found for those examinees who responded to the simplified items when compared to the regular items. Therefore, the differences in means seems to be due to a difference in the difficulty of the testlet, rather than in the version of items that were

contained in the testlet. This difference in the difficulty of the testlet was due to the fact that the testlets were constructed from raw items for which no prior item statistics were available.

LEP Examinees. A similar set of analysis of variance procedures and post hoc pairwise comparisons was computed for the LEP examinees. As expected, because of the very small and unequal cells, the overall F ratio at each of the two grade levels was not significant.

Reliability Coefficients for the 10 Item Testlets

Reliability coefficients were computed for each of the 10 item tests by form within grade, as shown on Table 4. Analysis of variance was used to compute alpha due to the presence of four short answer items scored 0, 1, 2. Algebraically alpha, which is derived by dividing the difference between the mean squares between people and the mean square due to residuals by the mean square between people, is identical to KR-20 (Hoyt, 1941).

As shown in Table 4, for the Non-LEP group, for a test of this length the coefficients were quite good. For the fourth grade sample, the coefficients for the regular and simplified items (Forms C and E) were .50 and .52, and (Forms D and F) .51 and .51. The corresponding coefficients for the sixth grade sample were (Forms C and E) .60 and .56 and (Forms D and F) .63 and .65. Thus, it would appear that the reliabilities of the two types of items do not differ.

TABLE 4
Reliability Coefficients For Non-LEP and LEP Examinees per 10 Item Testlet

Non-LEP	*N*	Grade	Form	Alpha
	1430	4	C	.50
	1426	4	D	.51
	1412	4	E	.52
	1416	4	F	.51
	1415	6	C	.60
	1416	6	D	.63
	1368	6	E	.56
	1423	6	F	.65
LEP	15	4	C	.19
	23	4	D	.23
	18	4	E	.55
	16	4	F	.00
	9	6	C	.00
	13	6	D	.75
	9	6	E	.02
	6	6	F	.63

For the very small LEP group, the coefficients confirmed that the testlets performed inconsistently with this group, with the result that testlets' reliability varied greatly under both conditions (simplified and regular items). For the fourth grade sample, the range was from .00 to .55; for the sixth grade sample, the range was from .00 to .75. For the regular condition, the range was .00 to .75 across the two grade levels. For the simplified condition, the range was .00 to .55 across the two grade levels. This variation is undoubtedly due to the small sample size.

Item Difficulty

Because each item was administered in both a regular and a simplified format, it is important to determine if there is any systematic difference in item difficulty by item format, and the magnitude of this difference. Where significant and substantial differences in item difficulty exist, it is also important to examine the items, in order to determine if there is an apparent reason for this difference. When the cause of such differences can be identified, this information can by used by test developers in future iterations of the test.

The procedure used to determine whether a significant difference exists between the item difficulty (P values) for each regular and simplified item requires the construction of 2x2 contingency tables to compute a Chi Square ($\chi 2$) for each pair of items. In the case of the short answer items, p-values were determined by summing the percent of examinees receiving either a 1 or a 2 on the item.

	Item 0	Item 1
Regular		
Simplified		

The procedure is equivalent to dividing the difference between two proportions by the standard error of the difference to obtain a normal deviate (z), which can then be referred to a table of areas under the normal curve to determine the level of significance. The chi square procedure is computationally convenient for testing the significance of the difference between two independent proportions. For one degree of freedom, $\chi 2$ is equal to the normal deviate squared. The data for each item for non-LEP and LEP student examinees on each form of the test is presented in Tables 5, 6, 7, and 8.

4th Grade Non-LEP Examinees. As shown in Table 5, when comparing the P values for form C (regular) to form E (simplified), 5 items (#1, 3, 5, 6, 10) had significantly higher P values in the regular format, 3 items (#2, 4, 9) had significantly higher P values in the simplified format, and 2 items were not significantly different (See Table 5). For form D (regular) and form F (simplified), no items were significantly different in their P values in the two formats. Clearly for the 4th grade non-LEP examinees, the simplified format was less likely to result in an easier item than the regular format.

TABLE 5
Comparison of Item Difficulty by Item Condition for 4th Grade Non-LEP Examinees

Item	Form C Regular P Values (n=1430)	Form E Simplified P Values (n=1412)	Significance	Item Type
1	.75	.59	.01	MC
2	.59	.66	.01	MC
3	.65	.61	.03	MC
4	.64	.74	.01	MC
5	.60	.52	.01	MC
6	.62	.50	.01	MC
7	.49	.52	.08	SA
8	.53	.56	.09	SA
9	.58	.62	.02	SA
10	.53	.40	.01	SA
	Form D Regular P Values (n=1426)	Form F Simplified P Values (n=1416)		
11	.68	.67	.55	MC
12	.55	.57	.26	MC
13	57	.57	1.00	MC
14	.73	.74	.61	MC
15	.50	.48	.26	MC
16	.57	.58	.88	MC
17	.47	.48	.55	SA
18	.61	.64	.11	SA
19	.63	.64	.56	SA
20	.43	.41	.29	SA

6th Grade Non-LEP Examinees. As shown in Table 6, for form C (regular) compared to form E (simplified), 2 items (#22, 29) had significantly higher P values in the regular format, 4 items (#24, 25, 26, 27) had significantly higher P values in the simplified format, and 4 items were not significantly different in their P values in the two formats. For form D (regular) and form F (simplified), 1 item (#31) had a significantly higher P value in the regular format, 3 items (#32, 35, 39) had significantly higher P values in the simplified format, and 6 items were not significantly different in their P values for the two formats. Thus, for the 6th grade students, neither format was likely to make a difference in item difficulty.

When comparing P values for the non-LEP groups, one must keep in mind that a very small difference in absolute value (.03) can produce a statistically significant difference at the $p<.05$ level when analyzing data based on large groups of examinees (See Table 6.).

TABLE 6
Comparison of Item Difficulty by Item Condition for 6th Grade Non-LEP Examinees

Item	Form C Regular P Values (n=1415)	Form E Simplified P Values (n=1368)	Significance	Item Type
21	.75	.74	1.00	MC
22	.39	.35	.04	MC
23	.67	.67	1.00	MC
24	.72	.81	.01	MC
25	.75	.80	.01	MC
26	.36	.40	.04	MC
27	.14	.19	.01	SA
28	.18	.17	.45	SA
29	.39	.27	.01	SA
30	.28	.30	.28	SA
	Form D Regular P Values (n=1416)	Form F Simplified P Values (n=1423)		
31	.75	.71	.02	MC
32	.77	.80	.05	MC
33	.80	.80	.93	MC
34	.77	.77	.89	MC
35	.62	.67	.01	MC
36	.33	.32	.52	MC
37	.60	.63	.11	SA
38	.73	.75	.16	SA
39	.36	.41	.01	SA
40	.21	.24	.11	SA

4th Grade LEP Examinees. As shown in Table 7, for form C (regular) and E (simplified), 1 item (#1) had a significantly higher P value in the regular format, 1 item (#9) had a significantly higher P value in the simplified format, and 8 items were not significantly different in their P values for the two formats. For forms D (regular) and F (simplified), 2 items (#18, 19) had significantly higher P values in the simplified format and 8 items were not significantly different in their P values for the two formats (See Table 7.).

TABLE 7
Comparison of Item Difficulty by Item Condition for 4th Grade LEP Examinees

Item	Form C Regular P Values (n=15)	Form E Simplified P Values (n=18)	Significance	Item Type
1	.73	.28	.02	MC
2	.67	.56	.72	MC
3	.53	.50	1.00	MC
4	.40	.50	.73	MC
5	.47	.44	1.00	MC
6	.40	.28	.49	MC
7	.27	.22	1.00	SA
8	.20	.50	.16	SA
9	.13	.50	.03	SA
10	.53	.17	.06	SA
	Form D Regular P Values (n=23)	Form F Simplified P Values (n=16)		
11	.57	.50	1.00	MC
12	.30	.38	1.00	MC
13	.39	.44	1.00	MC
14	.26	.44	.31	MC
15	.39	.19	.29	MC
16	.44	.38	.75	MC
17	.17	.19	1.00	SA
18	.35	.69	.05	SA
19	.22	.63	.02	SA
20	.13	.19	.67	SA

6th Grade LEP Examinees. As shown in Table 8, for forms C (regular) and E (simplified), 1 item (#21) had a significantly higher P value in the regular format and 9 items were not significantly different in their P values for the two formats. For forms D (regular) and F (simplified), no items were significantly different in their P values for the two formats (See Table 8).

TABLE 8
Comparison of Item Difficulty by Item Condition for 6th Grade LEP Examinees

Item	Form C Regular P Values (n=9)	Form E Simplified P Values (n=9)	Significance	Item Type
21	.78	.22	.03	MC
22	.33	.33	1.00	MC
23	.56	.33	.40	MC
24	.67	.44	.40	MC
25	.67	.33	.20	MC
26	.33	.33	1.00	MC
27	.11	.00	1.00	SA
28	.00	.11	1.00	SA
29	.33	.00	.21	SA
30	.00	.00	---	SA
	Form D Regular P Values (n=13)	Form F Simplified P Values (n=6)		
31	.46	.17	.33	MC
32	.46	.67	.63	MC
33	.54	.33	.63	MC
34	62	.33	.35	MC
35	31	.17	.63	MC
36	15	.00	.54	MC
37	31	.17	.63	SA
38	15	.17	1.00	SA
39	.23	.00	.52	SA
40	.00	.00	---	SA

Examination of Simplified Items

In cases where significant differences are found in the difficulty of test items, it is especially important to analyze the changes in wording that were made in the test items. In theory, the analysis of these changes will identify the features that make the item easier or more difficult. The features that cause differences in difficulty should show up most clearly in the items where the differences in difficulty are greatest. Thus, we examined the two items that showed the greatest difference in difficulty to see what might have caused the differences. It should be understood that for the Non-LEP group, all differences were small. Therefore, even the two most discrepant items show small differences in P values.

Grade 4, forms C and E, item 4. The difference in P values on this item was .10, in favor of the simplified version. (See Table 5.)

> For this item, the principal difference is that the task is clarified to the examinee. While the two versions of the item include a graphic that illustrates the principle tested by the item, only in the simplified version is the examinee told to look at the graphic. In the regular version, it is assumed that the examinee will look at the graphic. When the examinee's attention is called to the graphic, the answer becomes apparent to a greater number of examinees. It should be pointed out that this difference has nothing to do with simplified language. However, our team considered the lack of familiarity with American testing conventions when reviewing items. In cases where the task to be performed by the examinee was implicit, we often made the task explicit. In the case of this item, this may have helped some non-LEP students who otherwise would not have reacted to the graphic as the test developers assumed they would. In this case, the simplification process may have eliminated a weakness in the original item.

Grade 6, forms C and E, item 29. The difference in P values on this item was .12, in favor of the regular version. (See Table 6.)

> This item also contains a graphic, and the simplified version tells the examinee to look at the graphic. However, unlike the item in the example above, the simplified version is apparently more difficult. Perhaps the difference is due to other changes in the wording of the simplified version. The simplified version of item 29 avoids the use of the word "consequences" in order to keep the language simple. However, this word helps convey that the task is to identify the effect of the action introduced in the item. Also, in the simplified version a long stem in the form of an if...then...clause, is divided into two sentences. This may also have reduced the degree to which the item conveys that the examinee is to identify a causal relationship. Thus, at least for the fully English proficient student, "linguistic simplification" can inadvertently produce a more difficult item.

Summary and Discussion

When evaluating the efficiency of an accommodation, there are two issues to be determined. First, among those for whom it is not considered necessary, there is a need to understand whether it provides an unfair advantage to an examinee that receives it over one who does not. Second, if among the first group there is no advantage for those who receive it, then there is a need to understand whether the accommodation actually improves the performance of those who have special needs. One way to determine if an accommodation offers an unfair advantage, or whether it meaningfully assists students

with special needs, is through an experimental design whereby students are randomly assigned to treatments, with some students receiving the treatment and others not getting it. In this case, the treatment is the accommodation known as linguistic simplification.

In this study, the regular version of items on a 4th and 6th grade standards-based state assessment of Science was linguistically simplified by a team that included experienced test developers, applied linguists, and practicing science teachers. The process was done quickly and efficiently, without delaying the test development timelines. The simplified and regular version of items was then included on the portion reserved for field test items within forms of an operational state assessment. The forms were assembled so that the field test portions were identical except that two of the field test testlets consisted of regular items and two consisted of the same items in a simplified format. Thus, it was possible to compare the effects of linguistic simplification on test performance and item difficulty.

The tests were administered to large samples of 4th and 6th grade students participating in the Delaware Student Testing Program. Separate analyses of the results were completed for regular (non-LEP) and LEP students for each item condition (regular vs. simplified) and for each grade level. The results were broken down by total score on the testlet and by item difficulty (P value).

Only a small number of LEP students participated in this state assessment. Therefore, few significant differences were found in the LEP analyses, and it is not possible to draw any conclusions from the results regarding the effects of the simplified items on LEP students. However, the samples for the regular, fully English proficient students were quite large, with the result that conclusions can be drawn based on the data.

The results of the study support the conclusion that among fully English proficient students, linguistically simplified tests and items are normally of no help or almost no help to students taking a test. That is, linguistically simplified items function like eyeglasses. If you don't need eyeglasses to see clearly, then glasses do not improve your vision. On the other hand, if a person has deficient vision, then glasses will improve vision. Thus, glasses level the playing field for those who need them, so that everyone is able to see with an adequate degree of clarity, while not giving those who use glasses an advantage over those who don't.

In this study, there was essentially no difference in the mean raw scores of English speaking students who took simplified testlets and those who took the same testlet with regular wording. This is an important finding, because it shows that linguistic simplification can be used without fear of providing an unfair advantage to those who receive it, and thereby affecting the comparability of scores across examinees obtained under this condition. With this knowledge in hand, educational testing specialists, concerned with the identification of ways to meaningfully include more students in the assessment program, can offer linguistically simplified science assessments to limited English proficient students without fear of providing them with an unfair advantage on the test.

Other studies should now address the issue of the usefulness of linguistic simplification for LEP students taking formal and high stakes assessments. If experimental studies involving large samples of LEP students who are randomly assigned to treatments show that those LEP students who receive simplified items perform statistically and meaningfully better than those who receive the regular unsimplified version of such items, then the utility of linguistic simplification in meeting the needs of LEP test-takers will be established. Such studies will have to take place in states and large districts with large numbers of LEP students.

Another issue addressed by this study, although less formally, is the cause of differences in item difficulty when this occurs following linguistic simplification. A review of the items on which significant differences were found, did not show clear trends. This was probably due to the fact that for the fully English proficient student, the linguistic simplification did not help much, even when the differences were statistically significant. Again, should experimental studies involving large numbers of LEP students be conducted, it may be that differences will be considerably larger than those obtained here with non-LEP students. In that case, subsequent qualitative review of test items may produce a clearer understanding of the kinds of items that can benefit from linguistic simplification, and the kinds of revisions that are most successful in simplifying items for ELLs.

While it is unfortunate that the study did not identify and confirm the effectiveness of linguistic simplification for ELLs, the study was successful in showing that tests and items can be linguistically simplified without compromising score comparability. Of course, the process of linguistically simplifying test items must be carried out with care. The result of the process of linguistic simplification must be to make the item accessible to ELLs and not to alter the difficulty of the content. At times, language and content interact, and in these cases, it is not possible to linguistically simplify items without simplifying the content. Because the simplification process must be managed with caution, like item writing in general, it cannot be assumed that all linguistic simplification efforts will achieve the same result. However, if a future study demonstrates that linguistic simplification is effective for ELLs, additional research efforts will need to identify the linguistic features of items that cause problems for ELLs, and the procedures to be observed and the linguistic or organizational features to be implemented in the revision of test items. We encourage others to pursue this promising avenue for future research involving the testing of English language learners.

References

American Educational Research Association, American Psychological Association, National Council for Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

American Educational Research Association. (2000). Position statement of the American Educational Research Association concerning high-stakes testing in Pre-K-12 education. *Educational Researcher, 29(8),* 24-25.

Abedi, J., Kim-Boscardin, C., and Larson, H. (2000) *Summaries of research on the inclusion of students with disabilities & limited English proficient students*. CRESST: Los Angeles, CA.

Abedi, J. (1997). *Impact of selected background variables on students' NAEP math performance*. Los Angeles: Center for the Study of Evaluation. Draft deliverable.

American Institutes for Research (AIR). (1998, November). *Background paper reviewing laws and regulations, current practice, and research relevant to inclusion and accommodations for students with limited English proficiency*. Palo Alto, CA: Author.

Committee for Economic Development. (2001). *Measuring what matters: Using assessment and accountability to improve student learning*. Author: New York.

Delaware Department of Education, Assessment and Accountability Branch. (1999, January 4). *Delaware Student Testing Program: State summary report, 1998 administration*. Dover: Author. http://www.doe.state.de.us.

Delaware Department of Education, Assessment and Accountability Branch. (1999, March 31). *Guidelines for the inclusion of students with disabilities and students with limited English proficiency*. Dover: Author.

Duncan, D.B. (1955). Multiple range and multiple F tests. *Biometrics*, 11, 1-42.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, § 1111, 108 Stat. 3525 (1994).

LaCelle-Peterson, M. & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review*, 64(1), 55-75.

Miles, J. E., Rivera, C., & Stansfield, C.W. (2000). *Leveling the assessment 'playing field': Making science test items accessible to English language learners*. Arlington, VA: George Washington University, Center for Equity and Excellence in Education.

Rivera, C. & LaCelle-Peterson, M. (1993). *Will the national education goals improve the progress of English language learners? ERIC Digest.* Washington D.C.: ERIC Clearinghouse on Language and Linguistics. ED 362 073.

Rivera, C. & Vincent, C. (1997). High school graduation testing: Policies and practices in the assessment of English language learners. *Educational Assessment*, 4(4), 335-55.

Rivera, C. & Stansfield, C.W. (1998). Leveling the playing field for English language learners: Increasing participation in state and local assessments through accommodations. In R. Brandt, Editor, *Assessing Student Learning: New Rules, New Realities* (pp. 65-92). Arlington, VA: Educational Research Service, 1998.

Rivera, C., Stansfield, C.W., Scialdone, L., & Sharkey, M. (2000). A*n analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998-1999.* Arlington, VA: George Washington University, Center for Equity and Excellence in Education.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40, 87-104.

Shepard, L.A., Taylor, G.A., & Betebenner, D. (1998 September). *Inclusion of limited English proficient students in Rhode Island's grade 4 mathematics performance assessment. Center for the Study of Evaluation Technical Report No. 486.* Los Angeles: University of California at Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Stancavage, F. Allen, J., & Godlewski, C. (1996). Study of the exclusion and assessability of students with limited English proficiency in the 1994 Trial State Assessment of the National Assessment of Educational Progress. In *Quality and utility: The 1994 trial state assessment in reading.* Stanford, CA: National Academy of Education.

Teachers of English to Speakers of Other Languages Elementary and Secondary Education Act Reauthorization Task Force. (December 2000/ January/February 2001). Board endorses position papers for ESEA reauthorization effort. *TESOL Matters, 11(1)*, 1,4.

US Department of Education, Office of Elementary and Secondary Education (1996, October). *Title I, Part A Policy Guidance: Improving basic programs operated by local educational agencies, Guidance on standards, assessments, and accountability.* Washington, D.C.: Author.

U.S. Department of Education. (1999, November). *Peer reviewer guidance for evaluating evidence of final assessments under Title I of the Elementary and Secondary Education Act*. Washington, D.C.: U.S. Department of Education.

U.S. Department of Education (2000, July). *Summary guidance on the inclusion requirement for Title I final assessments*. Washington, D.C.: U.S Department of Education.

U.S. Department of Education, Office for Civil Rights. (2000). *The use of tests when making high stakes decisions for students: A resource guide for educators and policy makers*. Washington, D.C.: U.S. Department of Education.

Zlatos, B. (November 1994). Don't test, don't tell. *American School Board Journal*, pp. 24-28.